

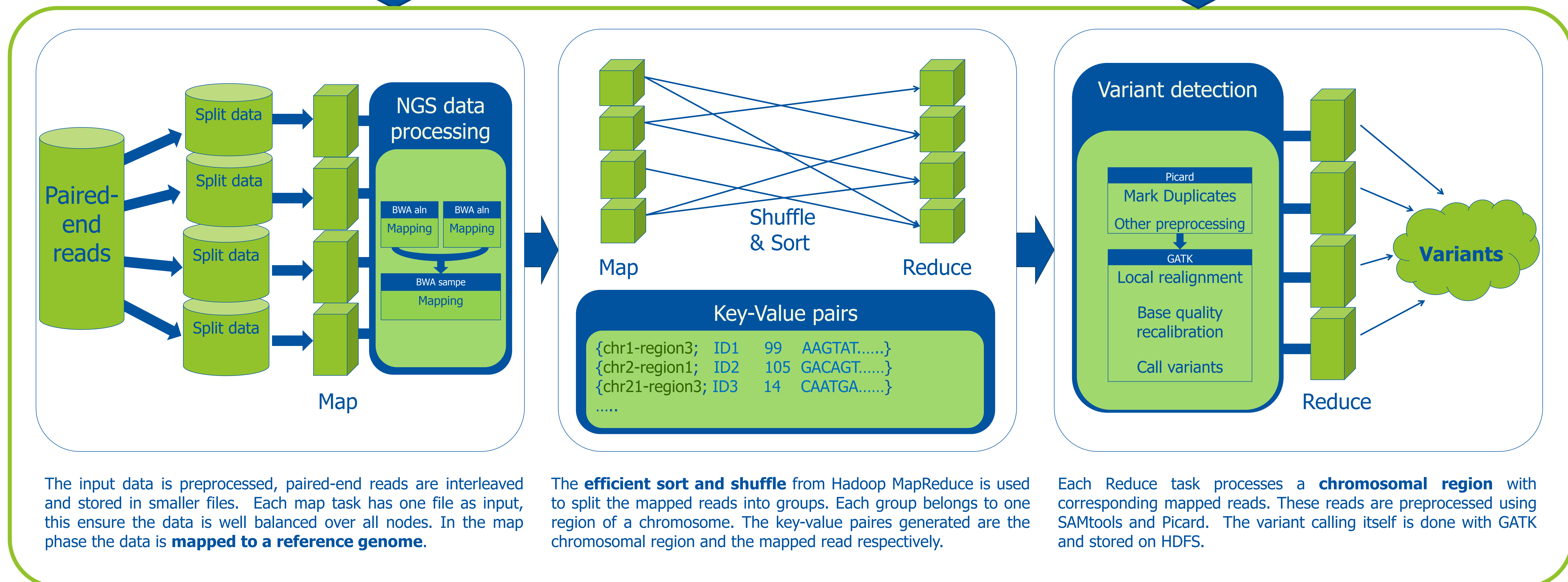
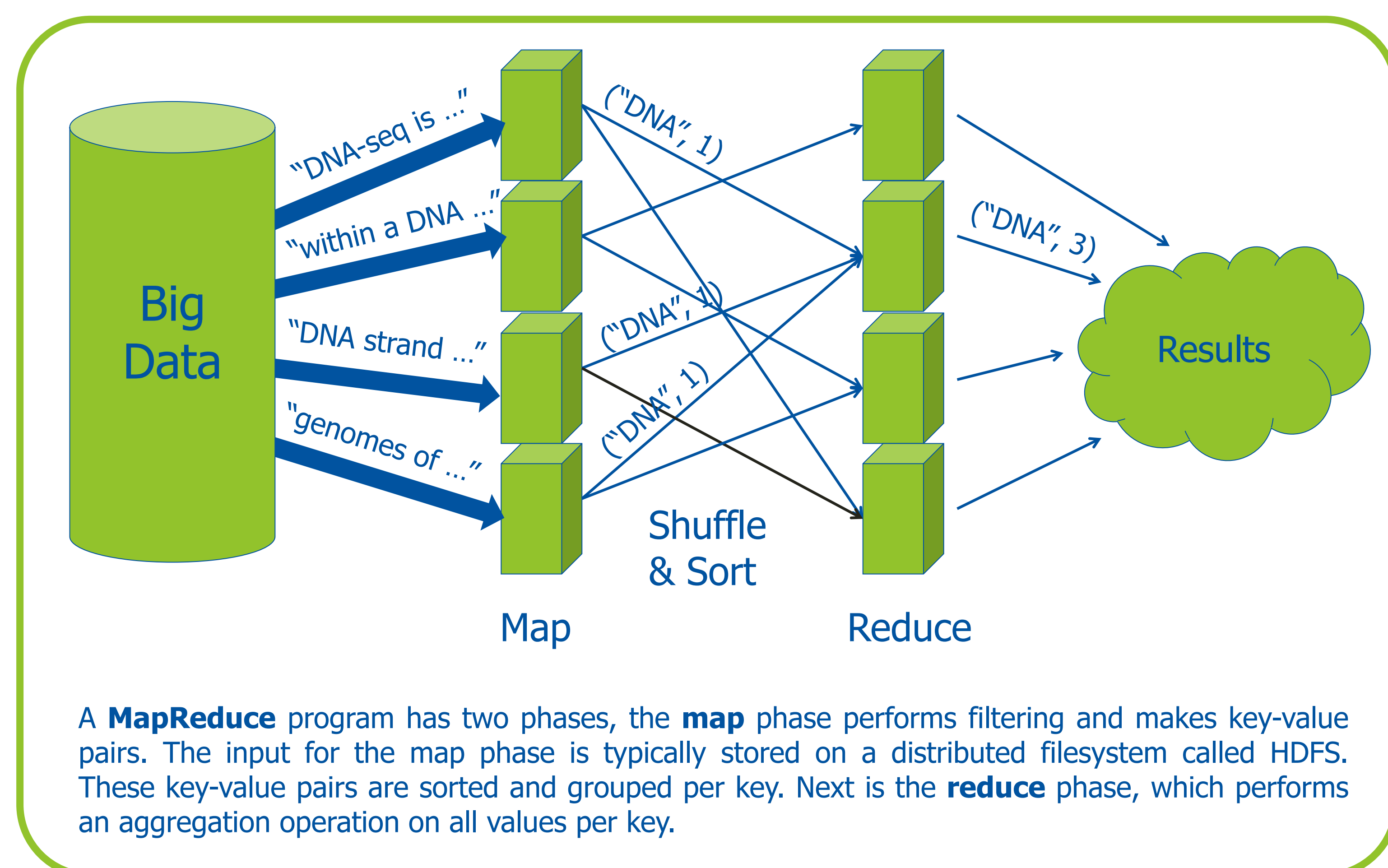
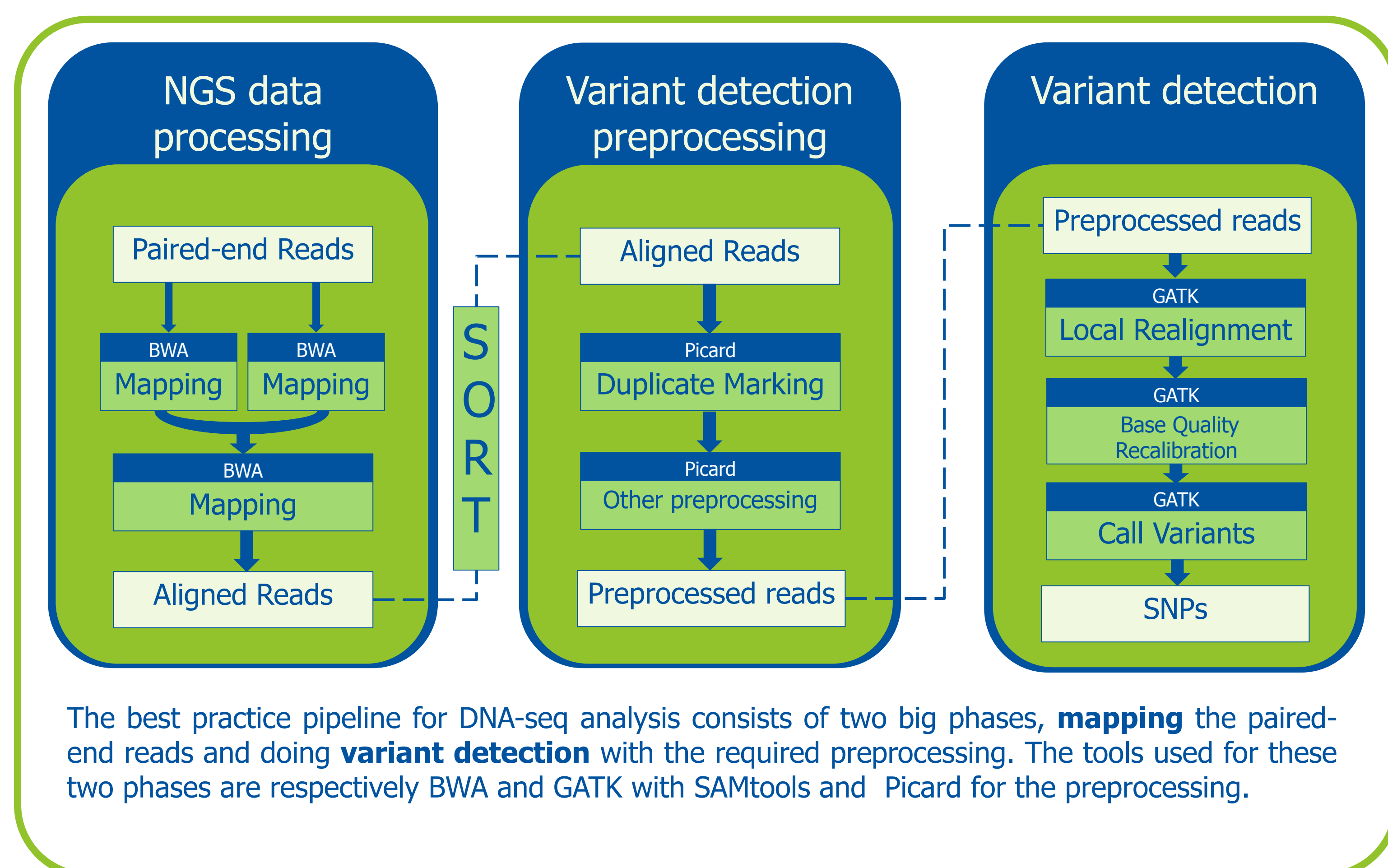
# Halvade

## whole genome analysis with MapReduce

Dries Decap<sup>1,5</sup>, Joke Reumers<sup>2,5</sup>, Charlotte Herzeel<sup>3,5</sup>, Pascal Costanza<sup>4,5</sup>, Jan Fostier<sup>1,5</sup>

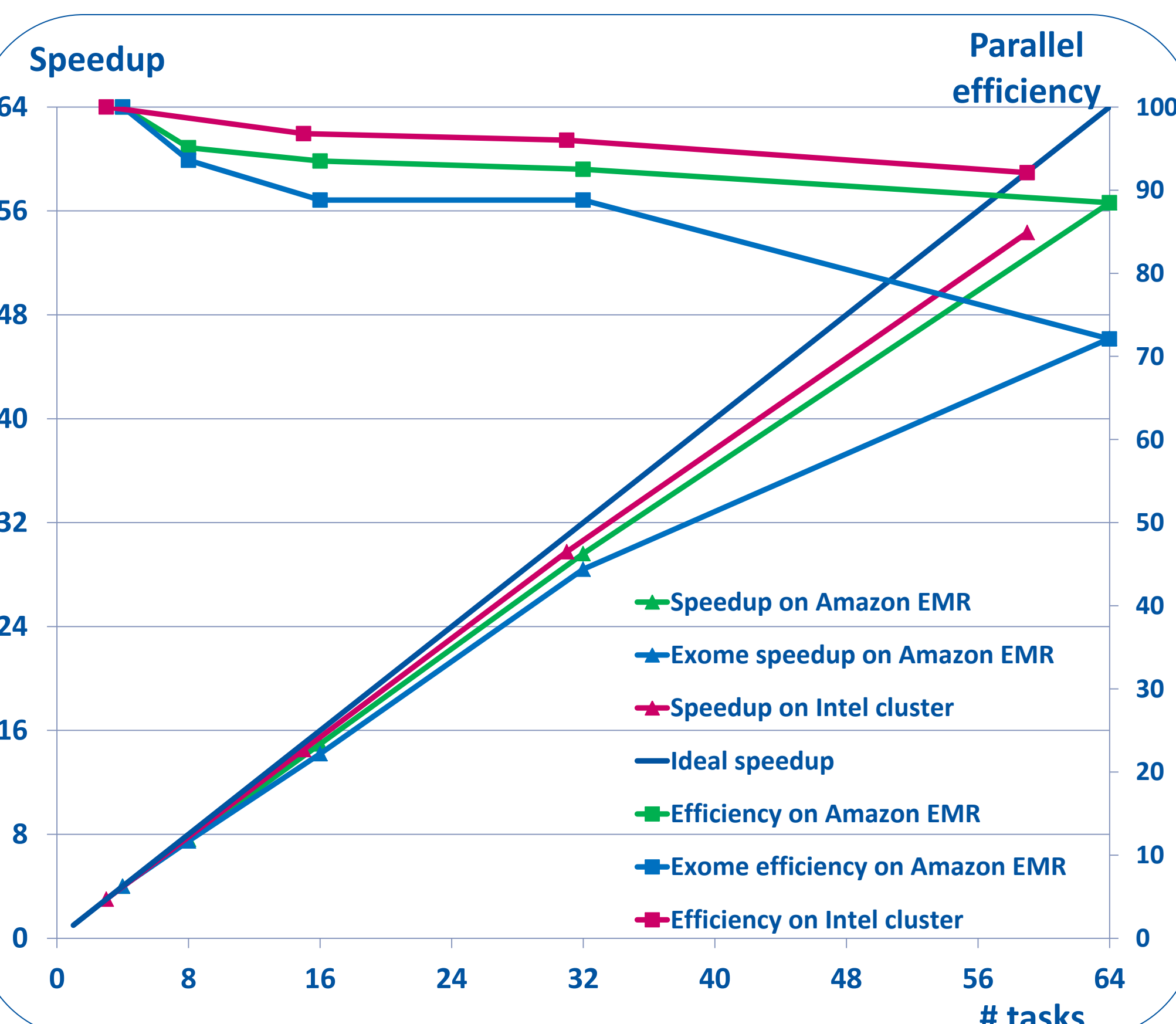
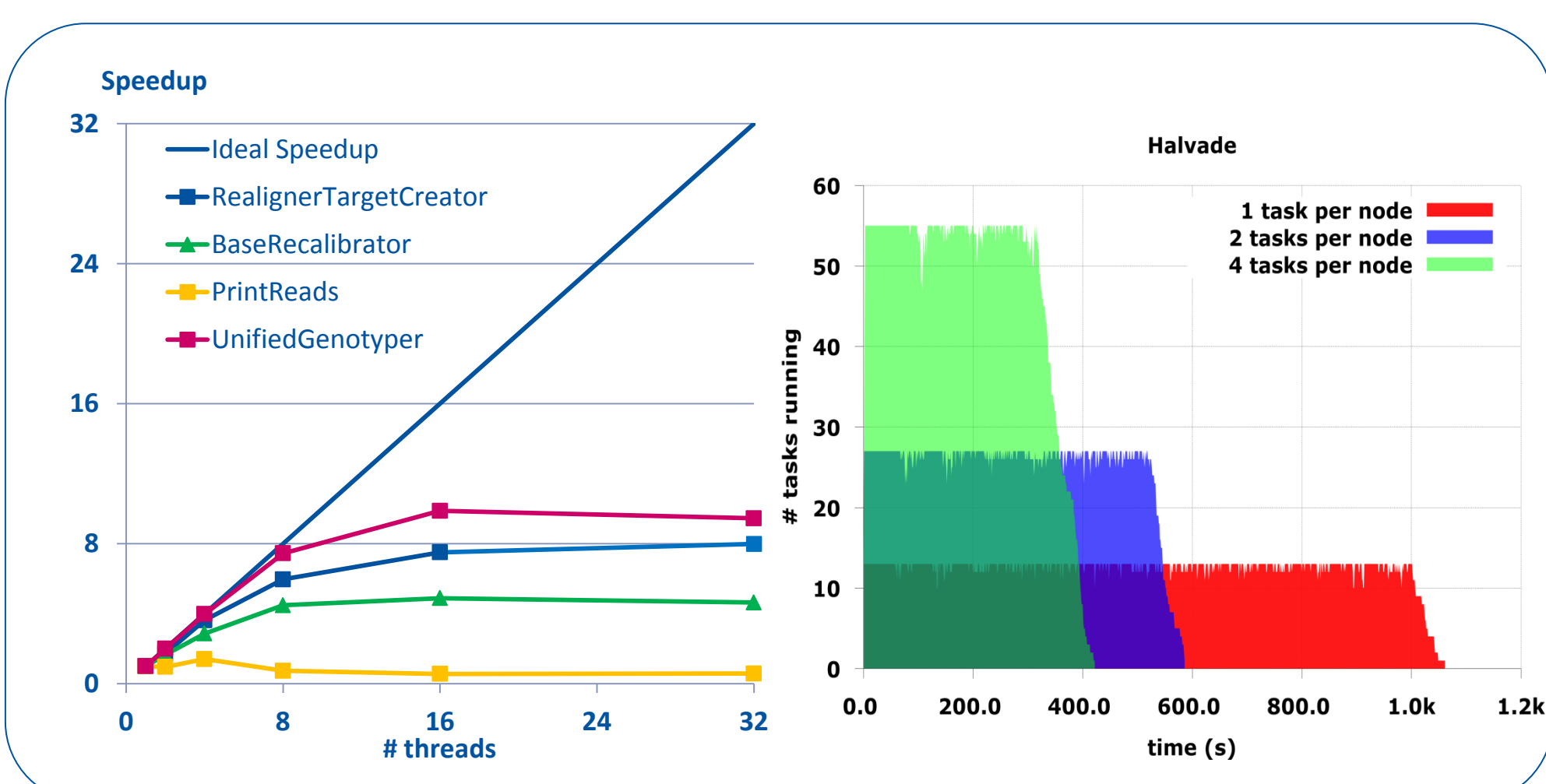
<sup>1</sup>: Ghent University – iMinds, <sup>2</sup>: JANSSEN R&D - Beerse (Belgium), <sup>3</sup>: IMEC, <sup>4</sup>: Intel Corporation NV/SA,

<sup>5</sup>: ExaScience Life Lab, Kapeldreef 75, B-3001 Leuven, Belgium



To utilize all available resources the individual tools are run in parallel. However some of the tools have limited parallel speedup (left graph) and cannot achieve optimal performance.

To reduce the absolute runtime, Halvade runs **multiple tasks simultaneously** on each node. The right graph shows the total number of map tasks running over time, which indicates that using multiple tasks per node causes a significant speedup.

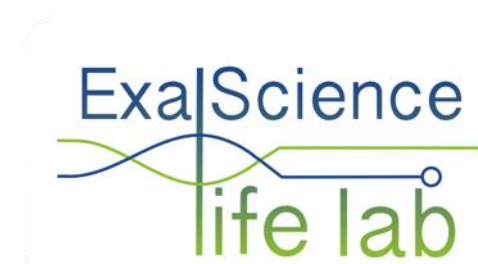


The parallel efficiency of Halvade was first assessed on the Intel Big Data cluster in Swindon, UK. This benchmark shows that Halvade has a parallel efficiency of 92,1% using 360 cores. In absolute runtimes, Halvade does **whole genome analysis in under 3 hours** using 15 nodes compared to 120 hours on one node without Halvade.

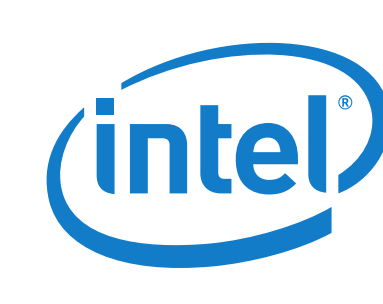
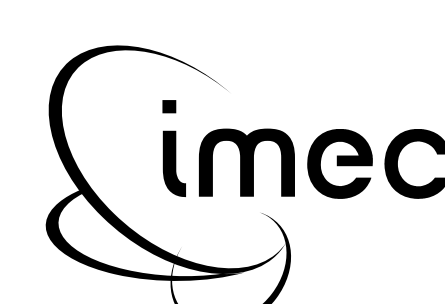
As a second assessment Halvade was run using Amazon EMR. This benchmark shows that, using 264 cores, Halvade achieves a parallel efficiency of 88,5%. In absolute runtimes this comes down to running whole genome analysis in under 3 hours and the **cost for whole genome analysis is ~111 USD**. Halvade supports exome analysis and gets the results in under one hour for ~20 USD on Amazon EMR.

To assess the accuracy of Halvade, the output was compared with a validation dataset. Halvade has an **accuracy of 99,4% for whole genome analysis** and 97,5% for exome analysis. Halvade can be used on any Hadoop MapReduce v2.0 or newer distribution including Cloudera and Amazon EMR and is freely available at <http://bioinformatics.intec.ugent.be/halvade>.

Dataset: NA12878, ~1,5 billion 100bp paired-end reads, 50x coverage, ~86GB compressed  
Intel Big Data Cluster: 15 nodes (dual socket Intel® Xeon® CPU E5-2695 v2 @ 2.40GHz, 62GB RAM)  
Amazon EMR: 16 nodes (32 vCPU, Intel® Xeon® CPU E5-2680 v2 @ 2.80GHz, 60GB RAM)



**Acknowledgments**  
This work is funded by Intel, Janssen Pharmaceutica, and Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT).



#### Notices and Disclaimers

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to <http://www.intel.com/performance>.